DOCUMENT RESUME

ED 048 352 TM 000 428

AUTHOR Whalen, Thomas E.
TITLE The Analysis of Essays by Computer: A Simulation of Teacher' Ratings.
PUB DATE Feb 71
NOTE 26p.; Paper presented at the Annual Meeting of the American Educational Research Association, New York, New York, February 4-7, 1971

EDRS PRICE EDRS Price MF-$0.65 HC-$3.29
DESCRIPTORS Comparative Analysis, *Computer Programs, *Essays, Grade 7, *Grade Prediction, Grades (Scholastic), Grading, Language Ability, *Models, Predictor Variables, *Writing Skills

ABSTRACT

The prediction of students' writing ability by the development of multiple regression models was investigated. Seventy-one essays and scores on the California Language Test by average seventh graders were used. Essays were entered in a modified Project Essay Grade (PEG) computer program. Results for three specific models are presented: (1) overall writing ability, (2) mechanical proficiency and (3) "standardized" language ability. Empirical cross-validations of these models resulted in predicted scores significant at the .01 level. The relative usefulness of twenty-six linguistic variables--seventeen used by researchers in earlier studies, with nine new ones added--was explored. Statistical procedures utilized in selecting subsets of these variables for predicting the above criteria are presented. Implications for future research in natural language analysis are discussed. (Author/LP)

THE ANALYSIS OF ESSAYS BY COMPUTER:

A SIMULATION OF TEACHER' RATINGS

by

Thomas E. Whalen

California State College, Hayward

THE ANALYSIS OF ESSAYS BY COMPUTER:

A SIMULATION OF TEACHER' RATINGS

Previous research (Page and Paulus, 1968) has shown the efficacy of grading student essays by computer. Using an actuarial approach, the researchers were able to obtain a multiple correlation of .72 between their thirty computer-derived predictors and an average of human judgments for five writing traits--content, organization, style, mechanics, and creativity. The sample used in this research consisted of essays written by students in grades eight through twelve.

In a related study by Janzen (1968), twenty-two linguistic variables similar to those used by Page and Paulus were utilized as criteria of writing ability. A factor analysis of scores on these variables resulted in five separate dimensions of writing (named by the author)--mechanical accuracy, fluency, sentence complexity, opionation, and emotionality. The author concluded that "this verifies Pages' findings that these variables are useful as 'proxes' for a measure of a student's ability to write English composition" (p. 47). This study utilized essays written by college students with a mean age of 18.8 years.

Two observations can be made concerning the above studies: (1) Both dealt with the analysis of fairly mature writing styles. Page's sample had an average I.Q. of about 114 and "could not be said to represent a random sample from the secondary population of the United States" (p. 20). (2) The one writing trait commonly agreed upon in the two studies was mechanical accuracy.

The purpose of this study was to seek answers to the following questions: (1) Do the linguistic variables used successfully by Page and Janzen serve equally well as predictors of writing ability for average seventh grade students? (2) Can the prediction of mechanical accuracy be improved through the use of additional variables and by achieving a more objective and reliable criterion measure of mechanics? and (3) Can computer evaluation of essays predict with any degree of success student scores on a standardized test of English mechanics including capitalization, punctuation, word usage, and spelling?

PROCEDURES

A sample of seventy-one essays written by average seventh graders was used in the study. Scores on the California Language Test, Junior High Level, for the same students were also utilized. The essays were entered as input via punched cards to a modified version of the Project Essay Grade (PEG) computer program (Paulus, 1969). Frequency counts for the following variables were made: Number of (1) paragraphs, (2) parentheses, (3) commas, (4) colons, (5) semicolons, (6) quotation marks, (7) question marks, (8) prepositions, (9) connective words, (10) spelling errors, (11) relative pronouns, (12) subordinating conjunctions, and (13) words on the Dale list. In addition, averages and standard deviations were calculated for word and sentence length. These measures were entered into a multiple regression equation to product overall essay grades as assigned by a panel of judges. The multiple correlation

and associated standard error of estimate were calculated and used as measures of predictive efficiency for the mathematical model.

A second equation was formed by adding six new stylistic variables and three new mechanics variables to the first equation. The two models were then compared to determine if the use of the new variables improved the predictability of essay quality.

Although prediction of overall essay quality was the chief focus in this part of the study, many minor hypotheses were also tested. For example, there was reason to believe that not all of the predictors would be correlated to overall quality in the same way for this sample of essays as for those of Page and Janzen. Because seventh graders have not yet mastered all of the basic writing skills, one might expect to detect excessive and incorrect use of such features as commas, colons, quotation marks, etc. Page found a high occurence of these features to be correlated positively with essay quality (p. 87). Correlations for these and other predictors were calculated to determine if their relationships with the criterion were stable across grade levels.

The present study attempted to improve on the predictability of mechanical accuracy in the following manner: Each essay was analyzed for mechanics alone by judges not invloved in the overall evaluation. All errors were tabulated and catagorized under sixteen previously established error-types. Prior research using this procedure (Whalen, 1969) has shown a high correlation between mechanical errors in essays and scores achieved by students on the California Language Test. In

addition, three new mechanics variables were added to the prediction equation--number of capital letters, number of capitalization errors, and number of usage errors. Since most errors in capitalization are errors of omission, it was hypothesized that a higher number of capital letters would correlate positively with mechanical accuracy. Capitalization errors were determined by checking each word in the essay against a dictionary of proper nouns. A word usage dictionary was also used. It contained 500 words and phrases commonly misused in student writing such as knowed, drownded, has went, could of, would of, etc. The prediction of mechanical accuracy was determined using the same linear regression technique as that used for overall quality.

The third question of interest involved the prediction of mechanical proficiency as measured by the California Language Test. Totals and subscores from the test including capitalization, punctuation, word usage, and spelling were used as the dependent variables. Separate regression equations were formed for each of these. Comparisons of multiple correlations and standard errors were made. Special concern was given to determining which proxes were most potent in the prediction of mechanical accuracy both for total test scores and for each of the four sub-categories.

RESULTS

Utility of Previously-Used Predictors

A question of initial interest was whether the variables used by

Page (1969) and Janzen (1969) at the high school and college levels could be utilized to predict essay grades for seventh grade writers. Therefore, a prediction model composed of seventeen variables used in common by those investigators was created. The multiple correlation for the seventeen predictors with the criterion essay grade was .85, an apparent improvement over the results of Page and Paulus. An estimate of the shrinkage in the mult-R from one sample to another was calculated using the Wherry formula (Kelly, 1947, p. 474). A shrunken coefficient of .75 was calculated. Since this mult-R was based upon a criterion of less than perfect reliability, it was attenuated (Thorndike and Hagen, 1957, p. 194) to compensate for the unreliability of the criterion. The reliability of scores for essays in the validation sample was calculated by analysis of variance (Ebel, 1951): It was .84. The attenuated mult-R was .82, a considerable improvement over the average adjusted mult-R of .74 for the Page and Paulus data (1968, p. 103).

Essay Grade

Natural curiosity provided an incentive for determining to what extent the nine new predictors could improve upon the prediction of essay grades. These new predictors included the three mechanics predictors discussed above plus six "stylistic" variables: (1) type-token ratio (the number of different words divided by total words), (2) occurrences of the word so, (3) occurrences of and, (4) occurrences of when, (5) occurrences of then, (6) occurrences of forms of the verb to be.

The new predictors were added to the previous model and a new mult-R was calculated. Results of the regression analysis are shown in Table 1.

The first variable selected was standard deviation of word length, indicating that none of the new variables correlated higher than .65 with the criterion. Occurrences of then proved to be an important indicator of writing quality for the sample data. These two variables alone generated a mult-R of .736 and accounted for approximately fifty-four per cent of the total variance (see Kelly, et. al., 1969, p. 66, or Darlington, 1968, p. 165).

Other new predictors which exhibited significant correlations with the criteria were capitalization errors and usage errors with coefficients of -.31 and -.35 respectively. However, in relation to the other predictors in the model, usage errors did not appear to contribute substantially toward the overall mult-R since it was entered at step twenty-five in the process.

The multiple correlation for all twenty-six variables was .905. This mult-R was adjusted for shrinkage and attenuation as before. The corrected mult-R was .84, an improvement of .02 over the previous model.

Mechanics

Results for the prediction of mechanical proficiency are presented in Table 2. Since the criterion was defined in terms of error scores, it is necessary to interpret predictor correlations with the criterion accordingly. Capitalization errors proved to be an important measure of mechanics for the sample data. This variable correlated highly with the total mechanics errors and was selected first by the computer

algorithm. It should be pointed out that the computer-derived frequency counts for capitalization errors were based upon a dictionary of proper nouns taken from the book Tom Sawyer, (all students reported on this novel) and were not based upon a universal proper noun list. Even so, it is well worth knowing that such a variable can be utilized quite effectively in situations calling for a restricted topic assignment.

Standard deviation of sentence length was selected second in the stepwise program. The bivariate correlation of .44 lent support to the hypothesis that a wide variation of sentence lengths, most probably due to the more frequent use of run-ons, is associated negatively with mechanical proficiency at the seventh grade level. Other important predictors included average word length, the number of capital letters, quotation marks, parentheses, and prepositions.

Three new predictors in addition to capitalization errors were included among the first eleven variables. The type-token ratio was selected sixth in the stepwise procedure; occurrences of then and usage errors were selected tenth and eleventh, respectively.

The mult-R for mechanics was .861. This was corrected for skrinkage to .63 using the Wherry formula. Since the reliability of the criterion measure could not be calculated the shrunken coefficient could not be corrected for attention.

Language Ability

The prediction of standardized test scores of English Mechanics

from a single sample essay was considered intuitively to be a most difficult task. Table 3 shows much better results than anticipated for this model. The first variable selected was average word length with a correlation of .50 with the criterion. Considering that word length has no direct relationship to mechanical proficiency, such a high relationship provides evidence of the robustness of this variable as a general predictor. Variable two in the model, common words on the Dale list, is an example of a suppressor variable at work. Though its correlation with the criterion was fairly low, it was highly related to average word length (-.45). The effect was to partial out a large portion of the residual error variance of average word length and thus make its relationship with language ability stronger.

Other predictors exhibiting a relatively high relationship with the criterion were standard deviation of sentence length, capitalization errors, occurrences of when, usage errors, number of prepositions, average sentence length, and standard deviation of word length. As was the case with the two previous models, not all of these variables were entered early in the computational process. The mult-R of .882, after correction for shrinkage and attenuation, was .72.

Forced Models

Because this study emphasized the development of efficient prediction models, additional techniques were employed to improve predictability. An attempt was made to reduce the number of variables to

the most parsimonious set of predictors possible.

One source of error in both the full and restricted models stemmed from the unreliability, or instability, of certain predictors. A technique for discovering the relative stability of predictors was developed. Essays in the developmental sample were randomly assigned to two sub-samples of twenty-four and twenty-three essays each. Correlation matrices for both sub-samples were generated. Comparisons were made between the correlations of predictors with the criteria for the two sub-samples. Predictor correlations with the criterion mechanics are exhibited in Table 4.

The z-values shown in Table 5 were calculated by using an equation given by Lordahl (1967, p. 273) for determining if two correlation coefficients have been sampled from the same population. Although it was already known that these coefficients were sampled from the same subpopulation, the z-values were useful in identifying those predictors whose relationship with the criterion was not stable. The absolute z-value is monotonically related to predictor stability; the lower the z-value, the more stable the predictor across sub-samples.

Variables twenty and twenty-four of Table 4, occurrences of and and number of capital letters, were characterized by high fluctuations from one sample to the other. And shifted from .40 to -.24, and number of capitals from .25 to -.36. Even variable three, average word length, fluctuated considerably, though its coefficient remained negative across samples. Three other predictor variables, number of colons, semicolons, and question marks, were found to have zero frequencies for the second

half of the developmental sample.

Additional evidence of the infrequent occurrence of several predictor measures is given in Table 5, which shows means and standard deviations for the twenty-six variables. Several variables including number of parentheses, colons, semicolons, question marks, and spelling errors had mean occurrence frequencies of less than .5. Moreover, their standard deviations were often considerably higher than their means, indicating highly skewed distributions. Although normal distribution of predictors is not a requisite assumption of linear regression analysis, there is little doubt that great that great departures from normality are bound to introduce additional error into the prediction models.

Using evidence gained from these data-analysis techniques, a limited subset of predictors was selected on the basis of their relative stability, their frequency of occurrence, and the magnitude of their correlations with the criteria. Tables 6 through 8 show the composition of three "forced" prediction models. It is important to note that, although the mult-R's for these models were lower than for the full models, the F-values were considerably higher.

The results of cross-validation of the forced models are shown in Table 9. Considerable improvement was noted for all models. Correlations between predicted and actual scores for essay grade, mechanics, and language ability were .60, .68, and .60, respectively. All of these coefficients are significant at the .01 level. Less success was noted for the California subtest models. However, the coefficients were all

significant at the .05 level.

DISCUSSION

Essay Grade Model

The final model constructed to predict essay grades was composed of only four variables--standard deviation of word length, occurrences of then, number of capitalization errors, and standard deviation of sentence length. These four variables were shown to have a multiple correlation with the criterion of .78, which, after cross-validation and correction for attenuation, was adjusted to .60, significant beyond the .01 level.

Apart from these statistical results, however, one might ask the question: How good are these results when compared with human judgments of writing quality? The following procedure was developed to answer that question. First, each of the five judge's ratings were compared with an average of ratings made by the other four judges. Secondly, the five correlation coefficients resulting from these comparisons were averaged (using Fisher's z-transformations). The resulting coefficient represented the average agreement between each judge and a combination of the other four judges. This procedure was carried out in exactly the same way using the set of scores generated by the computer. The computer scores were compared with pooled ratings from all combinations of four human judges. Correlations from these comparisons were then averaged, and the resultant coefficient was compared with the one representing

average agreement of the human judges. The results of this final comparison are shown in Table 10.

Clearly, the average human judge was able to do a somewhat better job in scoring the essays than the computer. However, just how important is the difference between .64 and .54 in terms of letter grades assigned to the essays? To answer this question, a procedure similar to the one above was used, except that numerical scores were converted back to their appropriate letter grade marks. In this case, it was determined that the average judge was successful in accurately scoring 10.6 of the twenty-four essays in the validation sample. This compared with ten correct predictions by the computer. Both the computer and the average judge missed one mark by two grade levels and the remaining marks by only one grade level. Since the five judges used in this study were selected on the basis of their agreement with known experts, it can be concluded that the computer model was about as successful in predicting essay grades as a well-qualified human judge.

Mechanics Model

The final model constructed to predict mechanical proficiency was composed of seven variables: (1) number of capitalization errors, (2) standard deviation of sentence length, (3) standard deviation of word length, (4) number of connectives, (5) occurrence of then, (6) average sentence length, and (7) number of usage errors. The variables in this model had a multiple correlation of .77 with the criterion. This

coefficient shrank to .68 after cross validation.

These results indicated that the mechanics model was the most reliable of the three major models in this study. This 7-variable model compared favorably with a 30-variable model constructed by Page and Paulus (1968). Those investigators reported a mult-R of .69 for the prediction of mechanics. This coefficient was statistically adjusted for shrinkage and then attenuated with a resultant mult-R of .69 (1968, p. 103).

In order to determine the general utility of the mechanics model, an attempt was made to predict overall essay grades by using the regression weights derived from the mechanics criterion. Surprisingly, the correlation between actual and predicted scores for essay grades was .60, an improvement of .05 above the results obtained from the essay grade model (disregarding correction for attentuation). Though this improvement could have occurred by chance, a possible explanation is that the regression weights computed for the mechanics model were more reliable due to the objective procedures used in defining the mechanics criterion. A further analysis indicated that predicted scores for the two models correlated .91 with one another with respect to the prediction of overall writing quality.

Language Ability Model

Two of the predictor variables in the 7-variable language ability model were different from those of the mechanics model. These were number

of subordinating conjunctions and occurrences of so. The other variables were standard deviation of word length, number of capitalization errors, occurrences of then, standard deviation of sentence length, and number of usage errors.

Although the mult-R of .73 for this model was somewhat lower than those for essay grade or mechanics, the adjusted coefficient of .60 indicated that in this study the machine prediction of total test scores for the California Language Test from a single essay was just as successful as the prediction of that same essay's overall letter grade. Similar success was not achieved for prediction of the subtests of the California Language Test. Although the models for capitalization, punctuation, usage, and spelling generated statistically significant scores, much less confidence should be placed in their reliability.

CONCLUSIONS

The following conclusions were reached on the basis of statistical analyses of the data:

1. In general, those variables used by previous investigators did appear to contribute substantially to the prediction of writing ability at the seventh grade level. In particular, the word and sentence length variables were more strongly related to writing quality at the seventh grade than for higher grade levels. A few of the predictors such as number of parentheses, colons, semicolons, and question marks exhibited

extremely low and erratic occurrence frequencies. Their appropriateness at the seventh grade level was doubtful. Two other variables, number of paragraphs and number of commas, were characterized by quite low correlations with the criteria. Apparently, the correct use of paragraphing and comma placement is not a well-established writing trait at the seventh grade level.

2. The use of additional variables in conjunction with an objectively-measured criterion of mechanical proficiency did appear to improve the predictability of this criterion. Several of the new variables including number of capitalization errors, occurrences of then, and number of usage errors were important contributors to the prediction of mechanical proficiency. One of these variables, occurrences of then, exhibited a highly negative relationship with both mechanics and overall writing quality. Thus, the use of this variable as a general stylistic predictor should be pursued. Other new predictors such as number of capital letters, occurrences of and, when, and the forms of to be were less successful. Occurrence frequencies for capital letters and the word and were erratic across essays. Perhaps, with a substantially longer sample of text, their use might be more profitable. Contrary to the pronouncements of rhetoricians, frequent use of forms of the verb to be does not appear to have a negative effect upon human judgment of writing quality or on mechanical proficiency. Probably, other factors far outweigh this feature in terms of what constitutes good writing by seventh grade students.

3. The machine prediction of language ability as measured by the

California Language Test was, indeed, a success. However, accurate prediction of four separate dimensions of the test was less successful. This was due, in part, to the lower reliability of the subtest scores. One variable which was expected to contribute strongly to this and the mechanics model was number of spelling errors. However, its relationship with both criteria was not especially strong. Although there were more than four spelling errors on the average in each essay, the computer was able to detect less than one of them. This would suggest that the dictionary of misspelled words should be augmented to include more words commonly misspelled by less sophisticated writers. Another dimension which was not properly represented in the model was punctuation. None of the punctuation variables appeared to contribute substantially toward predicting the criterion. If additional predictors can be found which adequately measure spelling and punctuation ability, the prediction of language ability should certainly be improved.

TABLE I

STEPWISE MULTIPLE REGRESSION: ESSAY GRADE

Step	Variable	r	b-wt	SE	t-Value	Mult-R	SE
1.	S.D. wd. lgth.	.65	12.19	12.84	0.94	.653	5.69
2.	"Then"	-.48	-1.29	1.07	-1.20	.736	5.14
3.	Subord. conj.	.15	1.31	0.61	2.14*	.772	4.88
4.	Prepositions	.39	0.13	0.22	0.59	.811	4.55
5.	Quest. Mks.	.17	4.33	6.06	0.71	.830	4.38
6.	Cap. errors	-.31	-0.52	0.60	-0.87	.840	4.32
7.	Paragraphs	-.07	-0.60	0.41	-1.44	.847	4.28
8.	"When"	-.16	-1.71	0.95	-1.79	.856	4.23
9.	Quotation mks.	.02	-0.35	0.80	-0.43	.864	4.16
10.	"To be"	.04	0.18	0.43	0.41	.869	4.15
11.	Commas	.01	-0.20	0.28	-0.74	.873	4.15
12.	Av. wd. lgth.	.60	21.15	17.18	1.23	.878	4.13
13.	Parentheses	.02	2.84	2.26	1.24	.881	4.15
14.	Connectives	-.04	-1.40	1.56	-0.89	.883	4.18
15.	Dale List	-.04	0.04	0.05	0.79	.885	4.21
16.	"So"	-.17	-0.77	0.85	-0.91	.887	4.25
17.	S.D. sent. lgth.	-.42	-0.46	0.41	-1.11	.888	4.30
18.	Av. sent. lgth.	-.44	0.47	0.50	0.94	.891	4.32
19.	Colons	-.19	-8.77	11.35	-0.77	.894	4.34
20.	No. of caps.	.11	-0.21	0.24	-0.89	.896	4.38
21.	Type-token	-.16	-47.13	46.57	-1.01	.902	4.34
22.	"And"	-.21	-0.21	0.36	-0.59	.904	4.39
23.	Spelling	-.19	-0.38	1.25	-0.30	.904	4.48
24.	Rel. pronouns	.09	-0.13	0.57	-0.24	.905	4.58
25.	Usage	-.35	0.12	1.23	0.10	.905	4.68
26.	Semicolons	-.02	0.24	3.38	0.07	.905	4.80

Intercept Constant = -62.95
F Mult-R = 3.468 (at step 26)
* Significant at .05 level

TABLE 2

STEPWISE MULTIPLE REGRESSION: MECHANICS

Step	Variable	r	b-wt	SE	t-Value	Mult-R	SE
1.	Cap. errors	.54	1.79	1.37	1.30	.537	12.10
2.	S.D. sent. lgth.	.44	0.42	0.95	0.44	.693	10.46
3.	Av. wd. lgth.	-.50	1.39	39.10	0.03	.732	10.00
4.	No. of caps	-.18	0.41	0.54	0.75	.747	9.87
5.	Quotation mks.	-.16	0.23	1.84	0.12	.770	9.58
6.	Type-token	.15	153.93	105.99	1.45	.786	9.40
7.	Parentheses	-.09	-5.82	5.16	-1.12	.803	9.19
8.	Ave. sent. lgth.	.37	0.55	1.14	0.48	.813	9.09
9.	Prepositions	-.37	-0.55	0.50	-1.10	.821	9.04
10.	"Then"	.23	2.41	2.44	0.98	.824	9.08
11.	Usage	.19	-5.23	2.88	-1.86	.826	9.16
12.	Colons	-.05	-34.95	25.83	-1.35	.829	9.23
13.	S.D. wd. lgth.	-.47	-42.88	29.23	-1.46	.832	9.29
14.	Subord. conj.	-.10	-1.65	1.39	-1.19	.837	9.31
15.	Spelling	.14	3.73	2.85	1.3[illegible]	.841	9.34
16.	Connectives	-.19	-4.12	3.56	-1.15	.846	9.37
17.	Commas	.11	-0.5[illegible]	0.64	-0.78	.852	9.35
18.	Paragraphs	-.02	0.72	0.94	0.76	.857	9.37
19.	"When"	-.16	1.27	2.17	0.58	.859	9.47
20.	Semicolons	-.02	-3.07	7.70	-0.39	.861	9.61
21.	Question mks.	.17	3.04	13.79	0.22	.861	9.79
22.	"To be"	.04	-0.22	0.99	-0.22	.861	9.98
23.	"And"	-.21	-0.11	0.83	-0.13	.861	10.19
24.	Rel. pronouns	.09	0.16	1.30	0.12	.861	10.42
25.	"So"	-.17	-0.02	1.94	-0.01	.861	10.66
26.	Dale List	-.04	0.00	0.11	0.01	.861	10.93

Intercept Constant = 16.38
F Mult-R = 2.214 (at step 26)

TABLE 3

STEPWISE MULTIPLE REGRESSION: LANGUAGE ABILITY

Step	Variable	r	b-wt	SE	t-Value	Mult-R	SE
1.	Av. wd. Lgth.	-.50	-68.86	38.42	-1.79	.504	13.09
2.	Dale list	-.18	-0.18	0.11	-1.67	.684	11.18
3.	Spelling	.24	0.72	2.80	0.26	.724	10.70
4.	Colons	-.06	24.16	25.38	-0.95	.750	10.37
5.	Subord. conj.	-.28	-1.62	1.36	-1.18	.772	10.09
6.	S.D. sent. lgth.	.37	-0.01	0.03	-0.02	.793	9.80
7.	"Then"	-.11	1.77	2.39	0.74	.800	9.76
8.	Cap. errors	.36	1.08	1.35	0.80	.808	9.72
9.	Rel. pronouns	-.14	-.83	1.28	0.65	.816	9.67
10.	Quotations	-.28	-1.27	1.81	-0.70	.824	9.59
11.	"When"	.35	1.22	2.13	0.57	.828	9.63
12.	Type-token	.21	231.74	104.13	2.22*	.833	9.65
13.	No. of caps	.24	1.12	0.53	2.09*	.855	9.17
14.	"So"	.17	2.47	1.90	1.29	.864	9.05
15.	"To Be"	-.09	0.80	0.98	0.82	.869	9.02
16.	Connectives	-.08	1.66	3.50	0.47	.873	9.07
17.	Usage	.40	1.33	2.75	0.48	.876	9.10
18.	Question mks.	-.11	-5.85	13.55	-0.43	.878	9.20
19.	Parentheses	.09	-2.18	5.07	-0.43	.879	9.32
20.	Prepositions	-.37	0.11	0.49	0.23	.880	9.46
21.	"And"	-.05	-0.34	0.82	-0.42	.881	9.62
22.	Av. sent. lgth.	.35	0.26	1.12	0.23	.881	9.81
23.	Semicolons	-.05	-0.87	7.56	-0.11	.881	10.02
24.	S.D. wd. lgth.	-.47	-3.62	28.72	-0.12	.881	10.24
25.	Commas	-.09	-0.05	0.63	-0.09	.881	10.48
26.	Paragraphs	-.17	-0.07	0.93	-0.08	.882	10.73

Intercept Constant = 185.73
F Mult-R = 2.681 (at step 26)
* Significant at .05 level

TABLE 4

CORRELATION OF PREDICTORS FROM FIRST AND SECOND HALVES OF THE DEVELOPMENTAL SAMPLE WITH MECHANICS ERRORS

	Predictors	First Half (N=24) r	Second Half (N=23) r	z-Value
1.	Av. Sent. lgth.	.18	.43	-.90
2.	S.D. sent lgth.	.46	.45	.06
3.	Av. wd. lgth.	-.14	-.64	1.98
4.	S.D. wd lgth.	-.15	-.60	1.73
5.	Type-token	-.09	.27	-1.18
6.	Paragraphs	-.34	.06	-1.31
7.	Parentheses	.11	-.18	.93
8.	Commas	-.08	-.14	.19
9.	Colons	-.09	0.0	-.29
10.	Semicolons	-.06	0.0	-.19
11.	Quotation mks.	.06	-.21	.86
12.	Question mks.	-.20	0.0	-.64
13.	Dale list	-.26	.08	-1.12
14.	Prepositions	-.29	-.40	.38
15.	Connectives	-.12	-.26	.48
16.	Subord. conj.	-.03	-.14	.35
17.	Spelling errors	.16	.22	-.19
18.	Rel. pronouns	-.27	-.18	-.32
19.	"So"	-.26	.19	-1.47
20.	"And"	.40	-.24	2.11
21.	"When"	.13	.07	.19
22.	"Then"	.40	.18	.77
23.	"To Be"	.08	-.21	.93
24.	No. of capitals	.25	-.36	2.05
25.	Cap. errors	.62	.54	.42
	Usage errors	-.04	.31	-1.15

TABLE 5

MEANS AND STANDARD DEVIATIONS OF PREDICTORS

	Predictors	Mean	Standard Deviation
1.	Av. sent. lgth.	15.53	6.22
2.	S.D. of sent. lgth.	7.80	6.01
3.	Av. wd lgth.	4.02	0.19
4.	S.D. of wd. lgth.	1.87	0.19
5.	Type-token	0.47	0.04
6.	No. of Paragraphs	2.85	2.16
7.	No. of parentheses	0.20	0.43
8.	No. of commas	5.10	5.39
9.	No. of colons	0.01	0.12
10.	No. of semicolons	0.06	0.29
11.	No. of quotes	0.77	1.57
12.	No. of quest. mks.	0.01	0.19
13.	Common wds. on Dale	146.90	23.92
14.	No. of prepositions	17.51	4.11
15.	No. of connectives	0.52	0.79
16.	No. of subord. conj.	4.17	2.40
17.	No. of spelling errors	0.41	1.01
18.	No. of rel. pronouns	1.51	1.79
19.	Occurrences of "so"	1.20	1.31
20.	Occurrences of "and"	7.70	3.07
21.	Occurrences of "when"	1.69	1.72
22.	Occurrences of "then"	1.00	1.31
23.	Occurrences of "to be"	8.14	2.75
24.	Number of capitals	32.55	10.61
25.	Capitalization errors	.77	1.56
26.	Usage errors	1.14	1.34

TABLE 6

FORCED ESSAY GRADE MODEL

	Variable	b-wt	t-Value	Mult-R	F-Value
1.	S.D. of wd. lgth.	19.41	3.86**	.65	33.45
2.	"Then"	-1.93	-3.50**	.74	25.99
3.	Cap. errors	-1.06	-2.30*	.77	20.51
4.	S.D. of sent. lgth.	-0.14	-1.20	.78	15.90**
	Intercept	-8.97			

* Significant at .05 level
** Significant at .01 level

TABLE 7

FORCED MECHANICS MODEL

	Variable	b-Wt	t-Value	Mult-R	F-Value
1.	Cap. errors	3.85	4.05**	.54	18.21
2.	S.D. sent. lgth.	1.48	2.87**	.69	20.38
3.	S.D. wd. lgth.	-27.89	-2.39*	.72	15.36
4.	Connectives	-3.17	-1.58	.73	12.16
5.	"Then"	2.19	1.89	.75	10.23
6.	Av. sent. lgth.	-0.93	-1.79	.76	9.17
7.	Usage errors	-1.79	-1.31	.77	8.25**
	Intercept	71.30			

* Significant at .05 level
** Significant at .01 level

TABLE 8

FORCED LANGUAGES ABILITY MODEL

Variable	b-Wt	t-Value	Mult-R	F-Value
1. S.D. Wd. lgth.	-16.10	-1.18	.47	12.67
2. Subord, conj.	-1.91	-2.62*	.59	11.89
3. Cap. errors	2.94	2.71**	.64	10.29
4. "Then"	2.88	2.24*	.68	9.20
5. S.D. of sent. lgth.	0.35	1.25	.70	8.05
6. "So"	2.35	1.42	.72	7.11
7. Usage errors	1.50	0.98	.73	6.23 **
Intercept	57.21			

* Significant at .05 level
** Significant at .01 level

TABLE 9

CROSS VALIDATION OF FORCED PREDICTION MODELS

Criteria	Mult-R	Shrunk.	Atten.
Essay Grade	.78	.55	.60 **
Mechanics	.77	.68 **	-
Language ability	.73	.58	.60 **
Cal. Capitalization	.57	.35 *	-
Cal. Punctuation	.63	.45 *	-
Cal. Usage	.56	.45 *	-
Cal Spelling	.66	.36	.40 *

* Significant at .05 level for one-tailed test.
** Significant at .01 level for one-tailed test.

TABLE 10

A COMPARISON OF COMPUTER RELIABILITY WITH THE AVERAGE RELIABILITY OF A SINGLE HUMAN JUDGE

Single Judge	Judge Reliability	Computer Reliability	Comparison Group
A	.67	.60	B,C,D,E
B	.63	.54	A,C,D,E
C	.62	.51	A,B,D,E
D	.68	.56	A,B,C,E
E	.58	.48	A,B,C,D

Average Judge Reliability = .64

Average Computer Reliability = .54

REFERENCES

Darlington, R.B. Multiple regression in psychological research and practice. Psychological Bulletin, 1968, 69 (3), 161-182.

Ebel, R.L. Estimation of reliability ratings. Psychometrika, 1951, 16 (4), 407-424.

Janzen, H.L. A study of written language ability. Unpublished Master's thesis. The University of Calgary, Alberta, Canada, 1968.

Kelly, J.K., Biggs, D.L., and McNeil, K.A. Research design in the behavioral sciences: multiple regression approach. Carbondale, Ill.: Southern Illinois University Press, 1969.

Longyear, C.R. The analysis of essays by computer: linguistic analysis. U.S. Department of Health, Education and Welfare, May, 1970.

Lordahl, D.S. Modern statistics for behavioral sciences. New York: The Ronald Press Co., 1967.

Mosier, C.I. The need and means of cross-validation. Educational and Psychological Measurement, 1951, 11, 5-11.

Page, E.B., and Paulus, D.H. The analysis of essays by computer. USOE Project Number 6-1318. April, 1968.

Paulus, D.H. PEGFOR, a FORTRAN program for the analysis of natural language. Storrs, Conn.: The University of Connecticut, 1969.

Thorndike, R. L., and Hagen, E. Measurement and evaluation in psychology and education (Second Edition). New York: Wiley and Sons, Inc., 1961.

Whalen, T.E. Total English equals writing competence. Research in the teaching of English, 1969, 3 (1), 52-61.

Whalen, T.E. A comparison of language factors in primary readers. The Reading Teacher, 1970, 23 (6), 563-570.

Wherry, R.J. Comparison of cross-validation with statistical inference of betas and multiple R from a single sample. Educational and Psychological Measurement, 1951, 11, 23-28.